J Digit Imaging (2012) 25:454–465
DOI 10.1007/s10278-011-9443-5

Automatic medical image annotation and keyword-based image retrieval using relevance feedback

Byoung Chul Ko · JiHyeon Lee · Jae-Yeal Nam

Published online: 23 December 2011

Abstract This paper presents novel multiple keywords annotation for medical images, keyword-based medical image retrieval, and relevance feedback method for image retrieval for enhancing image retrieval performance. For semantic keyword annotation, this study proposes a novel medical image classification method combining local wavelet-based center symmetric–local binary patterns with random forests. For keyword-based image retrieval, our retrieval system use the confidence score that is assigned to each annotated keyword by combining probabilities of random forests with predefined body relation graph. To overcome the limitation of keyword-based image retrieval, we combine our image retrieval system with relevance feedback mechanism based on visual feature and pattern classifier. Compared with other annotation and relevance feedback algorithms, the proposed method shows both improved annotation performance and accurate retrieval results.

Keywords Image annotation · Random forests · Confidence score · Body relation graph · Relevance feedback

Introduction

As the digitalized medical images are considerably increased, medical image retrieval is an important issue to assist effective and accurate patient diagnoses. For example, the medical doctors search database images using keywords or visual features to find interesting or important cases and to compare visually similar images with their diagnoses. Moreover for fields such as case-based reasoning or evidence-based medicine, there is a need for finding similar medical cases [1].

Because keyword can represent semantics of images, keyword-based image retrieval is the typical query method. However, as a large number of medical images are annotated manually by doctors and medical experts, manual annotation is a time consuming and tedious work. To overcome the difficulty of manual annotations, content-based image retrieval (CBIR) was proposed. However, because CBIR systems in medical images typically relied on visual properties of an image, such as color, texture, shape, etc., simple query-by-image may fail in situations where the visual features cannot adequately capture semantic concept of images. Therefore, keyword-based searches have been the dominating approach for managing medical image database in contrast to natural image [2]. However, automatic keyword annotation considering semantic concept of medical images is a difficult work. To overcome the limitations of manual annotation and provide semantic keywords to medical images, some recent researches [3–5] have proposed automatic keyword annotation based on image classification.

Villena-Roman et al. [3] has proposed automatic annotation method using medical image and simple classification algorithm based on decision table classifier with weighted relevance aggregation. From the confidence of each of 57 classes, the class with the highest relevance is considered to be the class of the image and the images are annotated with Image Retrieval in Medical Applications (IRMA) code.

Setia et al. [4] proposed hierarchical classification scheme to reduce the computational complexity of the

B. C. Ko (✉) · J. Lee · J.-Y. Nam
Shindang-Dong Dalseo-Gu, Dept. Of Computer Engineering,
Keimyung University,
Daegu 704-701, South Korea
e-mail: niceko@kmu.ac.kr

J. Lee
e-mail: sonhyleejh@kmu.ac.kr

J.-Y. Nam
e-mail: jynam@kmu.ac.kr

support vector machine (SVM) classifier by using local relational features and ImageCLEF2006 and 2007 medical database. Classification starts from the top, each classification chooses one of the two possible child nodes until a leaf node is reached and images are annotated as a class name of a leaf node and IRMA code.

Mueen et al. [5] proposed a multilevel automatic medical image annotation and retrieval via keywords method based on concept hierarchy or class hierarchy. To address the semantic annotation, SVM-based approach by support vectors at different semantic level is used.

Amaral et al. [6] used three different classification methods, flat, axis-wise, and position-wise method by modifying SVMs separately and made use of pair-wise majority voting between methods by simply summing strings in order to produce a final annotation.

Despite keyword-based image retrieval providing easier query interface and more accurate retrieval results than CBIR, query results of annotated keywords still is far from user's satisfaction because image annotation is performed based on image visual features. Therefore, the combination of relevance feedback mechanism with keyword and visual feature is more desirable to enhance image retrieval performance. In relevance feedback mechanism, the user only needs to mark which images he or she thinks are relevant to execute the query. By the user's feedback action, weights for similarity [2, 7] or parameters for learning methods [8–15] are readjusted.

One of the conventional relevance feedback approaches is feature reweighting [2, 7]. In approach, if the feature of relevant images has a low variance, it indicates that these relevant images are consistent in this feature and that the feature should be assigned a relatively high weight. Conversely, a high variance gives a relatively small weight.

Recently, Rahman et al. [2] proposed medical image-retrieval method using multiclass SVM (MSVM) and feature-weighting method. After image retrieval, the feature weights for similarity fusion are calculated by considering both the precision and the rank-order information of top retrieved relevant images. The weights are dynamically updated by the system for each individual search to produce effective results.

Xu et al. [7] proposed a hybrid relevance feedback system for shape-based retrieval of spine X-ray image by improving the feature reweighting method. To enhance the feedback performance, the short-term memory to store feedback history is developed and an automatic weight-updating scheme is developed to present the images on which it is best for the user to provide feedback. However, conventional reweighting method requires large sample data for statistical parameter learning.

Recently, another key issue in relevance feedback is the learning strategy. Especially, SVM is one of the most effective learning techniques used in relevance feedback as image classification. The aim of SVM is to create a classifier that separates the relevant and relevant images and generalizes well on unseen examples [8].

Bao et al. [9] proposed two sampling algorithms for SVM-based relevance feedback using medical images: positive nearest neighborhood sampling method and positive margin sampling algorithm, which can select informative images to feedback to user. Then, these adopt 10-level relevance measurement and soft SVM to reduce the distance between the user query concepts and the target query images.

Liu et al. [10] presented medical images retrieval using semisupervised learning based on SVM for relevance feedback. This paper also introduced an algorithm about defining two learners, both learners are retrained after every relevance feedback round, and then each of them gives every image in a rank.

Oh et al. [11] proposed a relevance feedback approach based on incremental learning with SVM regression using mammogram images. Also, the authors present a new local perturbation method to further improve the performance of the proposed relevance feedback system.

Wei and Li [12] proposed a learning method for relevance feedback, which utilizes probabilistic model to generalize the two-class problem and provide an estimate of probability of class membership. To build the probabilistic model, SVM is applied to classify the mammograms and then scale them to the probability of class membership.

However, one major problem of learning method based on SVM is the insufficiency of labeled examples especially the small number of irrelevant examples, which might bring great degradation to the performance of the trained classifier [8]. In addition, SVM is not suitable when a feature has high-dimensionality as a result of computational complexity [13].

On the other hand, MacArthur et al. [14] proposed a relevance feedback retrieval system by learning a decision tree to uncover a common thread between all images marked as relevant using computed tomography (CT) greyscale images of human lungs. This tree is then used as a model for inferring which of the unseen images the user would most likely desire.

Lakdashti and Ajorloo [15] proposed a relevance feedback retrieval system based on interactive genetic algorithm to reduce the semantic gap of the present medical image retrieval systems. This system learns the user's semantics using relevance feedback and stores them in system's rules using *n*-dimensional hypercubes. However, in reality, since the user is interacting with the machine in real time, the algorithm should be sufficiently fast and avoid, if possible, heavy computations over the whole dataset.

In this work, we propose three frameworks that assigning multiple keywords into images, keyword-based image

retrieval, and supporting relevance feedback to retrieval system for reducing semantic gap between the user and a retrieval system in real time.

To improve keyword annotation performance, this study first proposes a novel medical image classification method combining local wavelet-based center symmetric–local binary patterns (WCS–LBP) with random forests. Second, for semantic keyword-based image retrieval, we propose confidence score assigning method to each annotated keyword by combining probabilities of random forests with predefined body relation graph. After confidence score assigning, we prove that our keywords having different confidence scores produce more efficient retrieval results when our method is applied image retrieval system. In addition, to overcome the limitation of keyword-based image retrieval, we combine keyword-based image retrieval with relevance feedback mechanism based on visual feature and pattern classifier. In relevance feedback mechanism, we also use the local WCS–LBP feature and random forest for confidence score updating.

Image database for annotation and retrieval task

This study used the ImageCLEF2007 [16] for the medical annotation and retrieval task. The main purpose of ImageCLEF is to provide a resource for benchmarking content-based image classification systems focusing on medical images [17]. ImageCLEF2007 consists of 10,000 training images and 1,000 test images from a total of 116 unique classes. The images included in ImageCLEF are annotated with complete IRMA code. The aim of the radiograph annotation task is to find out how well current image analysis techniques can identify image modality, body orientation, body region, and biological system examined based on the image content [4].

The IRMA code consists in four independent axes describing different content within the image [6, 17].

- the technical code (T) describes the image modality
- the directional code (D) models body orientations
- the anatomical code (A) refers to the body regions examined
- the biological code (B) describes the biological system examined.

According to the IRMA code, an image within the ImageCLEF 2007 database has a string of 13 characters, e.g., IRMA: TTTT-DDD-AAA-BBB. In this research, we use 900 training data and 1,500 test data from 30 frequently used classes.

The remainder of this paper is organized as follows. The "Image classification using local WCS–LBP and random forests classifier" section describes image classification method using local WCS–LBP and random forest. The "Automatic keyword annotation and confidence score assigning" section introduces keyword annotation and confidence score assigning method. In the "Keyword-based image retrieval and relevance feedback based on RF" section, keyword-based image retrieval and relevance feedback using random forests is introduced. The "Experimental results" section evaluates the accuracy and applicability of the proposed annotation and relevance feedback method based on experiments, and in the "Conclusion" section, we present some final conclusions and areas for future work.

Image classification using local WCS–LBP and random forests classifier

Local wavelet-based center symmetric LBP

Because LBP [18] describes gray scale local texture of the image with low computational complexity by using a simple method, it has been widely used in various computer vision applications. The original LBP descriptor forms different patterns based on the number of pixels by thresholding a specific range of neighbor sets with the center gray-scale intensity value. However, because the main problem of LBP is long feature dimension, Heikkilä et al. [19] proposed center symmetric LBP (CS-LBP). CS-LBP uses a modified scheme of comparing neighboring pixels of the original LBP to simplify the feature dimension from 256 to 16, while keeping the characteristics such as tolerance against illumination changes and robustness against monotonic gray-level changes.

In this paper, we use local wavelet based CS-LBP (WCS–LBP) [13] for feature extraction. Local WCS–LBP extracts a CS-LBP from all multiscale sub-images, including low-pass-filtered sub-images, after two-level wavelet decomposition. In general, since the X-ray image has strong edge distribution in the horizontal, vertical, and diagonal directions, the three high-pass-filtered sub-images (LH, HL, HH) have important properties when classifying image categories.

Seven sub-images are extracted after the two-level wavelet transform of an image, and all high-pass-filtered sub-images of each level are linearly combined as one wavelet energy of level 1, W_H^1 and level 2, W_H^2. The lowpass filtered subimage W_{LL}^2 is used by itself.

In addition, the major problems of medical images, especially radiograph images are high overlapping between image classes (i.e., hand is connected with carpal joint), we divide each sub-image into 4×4 local grids, and extract 16 dimensional local wavelet CS-LBPs from each sub-image.

The final histogram for each sub-image is generated by concatenating the local histograms. Since there are 16 sub-regions, the final dimension of the local WCS–LBP

histogram is 768 [(16×3)×16 subregions]. Finally, we concatenate all of the histograms to create the final local WCS–LBP histogram, as shown in Fig. 1. The concatenated final local WCS–LBP histogram is normalized to unit length using the Gaussian normalization method.

Random forest for image classification

For image classification, MSVM are reasonable choice due to its high performance and accuracy. However, MSVM are not suitable when the feature has high dimensionality and the database contains over 1,000 images, due to computational complexity. Therefore, the high dimensional local WCS–LBP feature vector with 768 dimensions might make training tasks very time consuming. According to the experimental results of Ko et al. [13], the processing speed for training and testing of the random forests (RF) is approximately 36.8 and 66 times faster than the MSVM method using the same training and testing images with the same local WCS–LBP feature vector.

In this paper, we have chosen to classify images using RF as proposed by Breiman [20]. This classifier has been shown to be effective in a large variety of high-dimensional problems with high computational performance and accuracy.

RF is an ensemble classifier of a number of decision trees, with each tree grown using some types of randomization. RF has a capacity for processing huge amounts of data with high training speeds, based on a decision tree. The structure of each tree in the RF is binary and is created in a top-down manner, as shown in Fig. 2.

In the training procedure, the random forest starts by choosing a random subset I' from the local WCS–LBP training data, I. At the node n, the training data I_n is iteratively split into left and right subsets I_l and I_r by using the threshold, t, and split function, $f(v_i)$, for the feature vector, v, using Eq. 1.

$$I_l = \{i \in I_n | f(vi) < t\},$$
$$I_r = I_n \backslash I_l. \tag{1}$$

Then, several candidates are randomly created by the split function and threshold at the split node. Among those, the candidate that maximizes the information gain about the corresponding node is selected. Consequently, a leaf node has a posterior probability and the class distributions, $p(c \mid n)$, are estimated empirically as a histogram of the class labels, c_i, of the training examples, i, that reached node n.

When classifying the test image, the local WCS–LBP histogram of the test image is created over the whole wavelet transform. The test image is used as input to the trained RF. The final class distribution is generated by ensemble (arithmetic averaging) of each distribution of all trees $L=(l_1, l_2,\ldots, l_T)$ and we choose c_i as the final class (f) of an input image if the final class distribution $p(c_i \mid L)$ has the maximum value.

$$f = \arg\ \max\left\{\frac{1}{T}\sum_{t=1}^{T} P(c_i|l_t)\right\} \tag{2}$$

The important parameters of RF are the depth of tree and the number trees, T. In this paper, we set a maximum tree depth to 20 and number of tree sets to 120 according to the experimental results of Ko et al. [13].

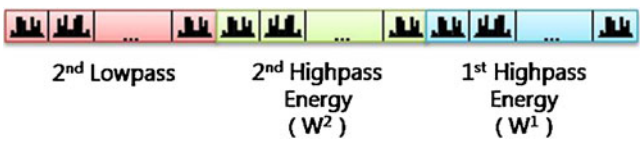

Fig. 1 Representation of the final local WCS–LBPs histogram generation. Local WCS–LBPs histograms are generated from one low-pass-filtered sub-image and the other two wavelet energies. All histograms are then concatenated to create the final local WCS–LBPs histogram

Automatic keyword annotation and confidence score assigning

Keyword annotation using image classification and body relation graph

There are three different classification strategies for annotating the radiographs as follows [4]:

- Flat strategy: an image is classified into one of the N base classes regardless of IRMA code. This strategy is usually performed using multiclass classifiers (e.g., MSVM).
- Axis-wise strategy: according to the IRMA code, four different multiclass classifiers are trained for individual technical, directional, anatomical and biological axes.
- Binary tree strategy: the classification tree is generated using agglomerative clustering and the class distances. Classification with binary classifier (e.g., SVM) starts from top, each classification chooses one of the two possible child nodes until a leaf node is reached.

In this paper, we use flat and axis-wise method concurrently: basically, we use the flat strategy for image classification using RF because it is simple and produce the best results out of the three different classification strategies as well as it requires N classifiers for N classes [4]. After RF is trained with training data collected from each axis, each test image is classified one of total classes. Then, an image is given the final IRMA code using the proposed body relation graph (BRG) that is the modification of axis-wise strategy. The BRG is designed for hierarchical representation of IRMA code. In BRG structure, BRG is composed of four

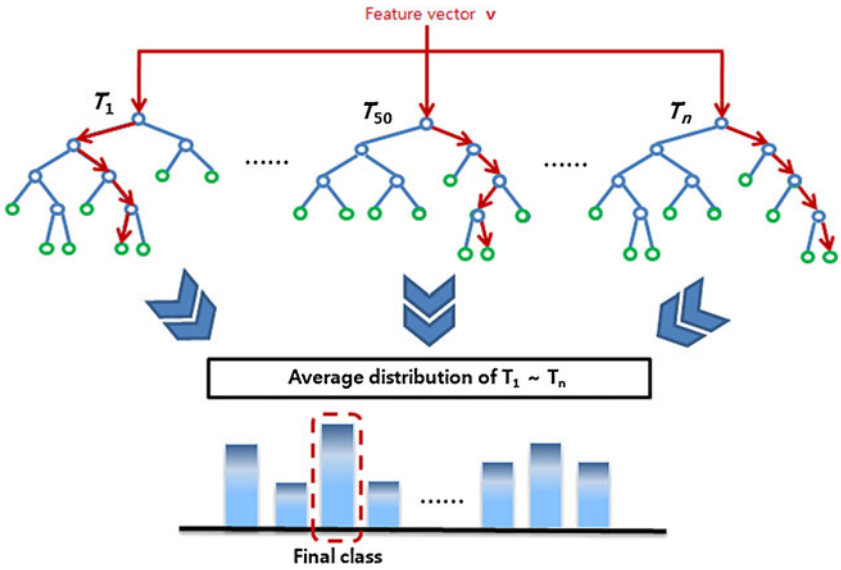

Fig. 2 Classification process using local WCS–LBPs with trained random forests. After training, the test image is classified into one class that has a maximum posterior probability

layers as it was defined in axis-wise strategy as described in Fig. 3: one top technical layer (T), three anatomical layers (A), and one biological layer (B). Especially, A layer is designed to consider human body parts and their relations. However axis-wise strategy should have four different multiclass classifiers for each individual axis, e.g., *N* classifiers for each layer, but, in BRG code, we need only major 30 classifiers and other 20 classes are classified according to the hierarchical connection with the 30 major classes. For example, if one image is classified class "Finger", it is also a member of higher classes "Hand", "Arm", and "Radiography" according to the BRG. Therefore, we can classify 50 classes using only 30 classifiers. As shown in BRG, the numbers marked in individual class is used for IRMA "D" and "A" code and other codes (T, B) are assigned automatically based on BRG.

For example, if a sample image is classified into "nose area", it has a code "213" for A layer as top down manner because "nose area (3)" is connected "facial cranium (1)", "cranium (2)". It also has "400" code for "D" because it is a member of "other orientation". Moreover, it has "1211" code for T and "700" code for B, automatically. Therefore, the final code for this image is "1211-400-213-700" and it has the following keywords according to the IRMA codes (Fig. 4).

The meaning of each class and the number of test images are shown in Table 1. As shown in Fig. 3 and Table 1, all classes for flat strategy are leaf nodes of BRG. Other classes can be predicted using BRG.

Confidence score assigning for image retrieval

After keyword annotation, an image has the at least eight to ten keywords. However, since all keywords have the same priority, if the user input a query "Left elbow", the retrieval system gives the all images which have the same keyword according to their sequential order regardless of their similarity. Therefore, we introduce the confidence score assigning method by giving the different priority to keywords by combining probabilities of RF and distance of BRG.

After one image is classified as one of 30 classes by RF, all classes which have the link with a classified class are given the relative distance dis based on the predefined BRG. The relative distance dis has a simple real number ($dis \geq 0$) to boost the relative difference between the major 30 classes and the connected higher classes. In this paper, we set difference of dis between major and linked classes as 0.8 because the dis shows similar detection results when it is $0.5 < dis \leq 1.0$ during experimentation. Therefore, the relative distance of each major classified class is 0 value and the relative distance of linked classes are increased plus 0.8 value as apart from a classified class. For example, if one image is classified as "nose area", dis for "nose area" has a 0 value and its higher node "facial cranium" has 0.8 value, "cranium" has 1.6 value, and "radiography" has 2.4 value.

After that, the relation weight (τ) of a classified class i and its linked parent classes are estimated using following scaled Fermi function [21];

$$\tau_i = \frac{2}{1 + \exp(\gamma \times dis_i)} \tag{3}$$

where, γ ($\gamma > 0$) is the control parameter and a small value of γ gives even large τ values, whereas a large value of γ make a large difference in maximal and minimal τ values. In this paper, because a good choice of adapting γ is to make τ over the medium value 0.5 [21], we set γ as 0.5.

The relation weight τ is maximal ($\tau = 1$) when the dis is 0 and tends to minimal ($\tau = 0$) for large dis for $dis \rightarrow \infty$. Therefore the influence of very small and very large dis is attenuated using Fermi function. By using Eq. 3, major 30

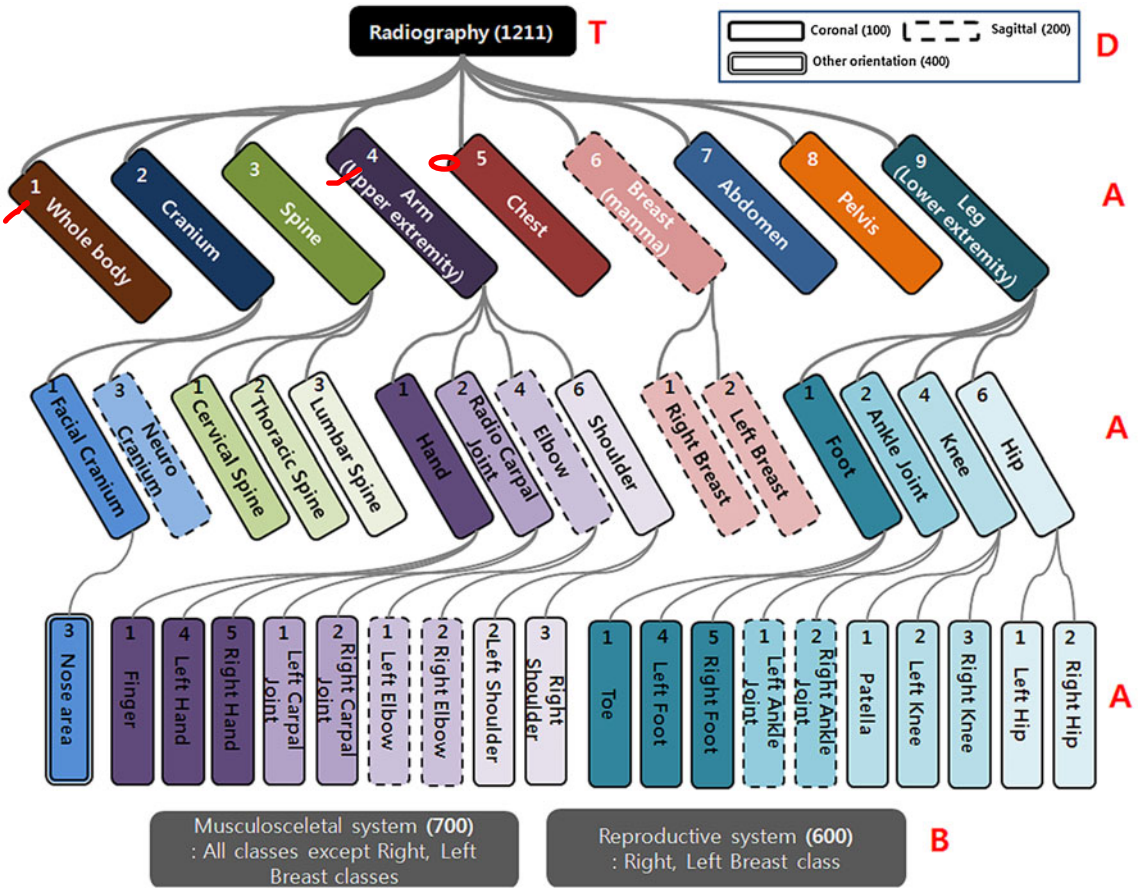

Fig. 3 Body relation graph (BRG) for efficient IRMA code assignment. The number of the box represents the IRMA code for the T, D, A, and B layer

classes have the relatively larger relation weight (τ) than other classes.

Then, confidence score Cf_i of a classified class i is estimated by using linear combination of the relation weight (τ_i) and its final class distribution $p(c_i \mid L)$ estimated from RF.

$$Cf_i = p(c_i|L) + \tau_i \tag{4}$$

From Eqs. 3 and 4, 30 keywords of major 30 classes have the larger confidence scores than other keywords of higher hierarchical classes. In contrast, confidence scores for keywords of layer D and layer B have the half confidence score of the lowest confidence score of layer A because D and B layer have the lower priority than A layer because they do not have relative relations with other classes.

After confidence scores on all connected classes are estimated, image has the multiple keywords according to the final IRMA code and all annotated keywords have different confidence scores based on their hierarchical relation.

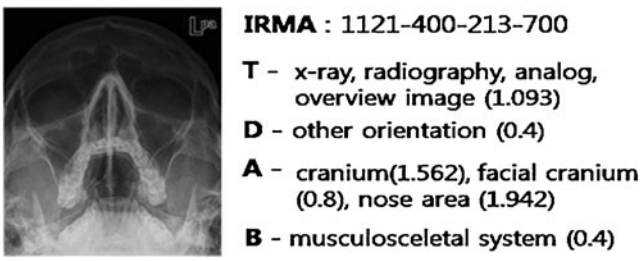

Fig. 4 Final IRMA code and its TDAB keywords. Confidence scores associated with each keyword are assigned within a bracket

Keyword-based image retrieval and relevance feedback based on RF

After keywords are annotated, the user input a query keyword and the system finds images including the same keyword in their annotation. In this system, we use a two-pass approach. First, our system retrieves a set of images having the same keyword with query q using linear search method and second, a more sophisticated matching technique is used: once the system finds a sufficient number of candidate images, the final distance is estimated by sorting the confidence values of keyword q included in candidate images and

Table 1 Description of the 30 classes for the flat strategy and the number of test images in each class

Class name	No. of images
Facial cranium	50
Cervical spine	50
Thoracic spine	50
Lumbar spine	50
Finger	50
Left hand	50
Right hand	50
Left carpal joint	50
Right carpal joint	50
Left shoulder	50
Right shoulder	50
Chest	50
Abdomen	50
Pelvis	50
Toe	50
Left foot	50
Right foot	50
Patella	50
Left knee	50
Right knee	50
Left hip	50
Right hip	50
Neuro cranium	50
Left elbow	50
Right elbow	50
Right breast	50
Left breast	50
Left ankle joint	50
Right ankle joint	50
Nose area	50

the top k nearest images are displayed in descending order of the final distance.

Even though keyword-based image retrieval provides accurate retrieval results, query results based on annotated keywords still is far from user's satisfaction. Therefore, the combination of relevance feedback mechanism with keyword and visual feature is more desirable to enhance image retrieval performance. In our relevance feedback mechanism, our objective is to retrain the decision trees of RF that have only two classes, relevant and irrelevant for next retrieval.

In our scenario, the user selects N relevant and irrelevant images after first image retrieval. Then, let $\mathrm{I} = \{(x_i, y_i)\}_{i=1}^{N}$ be the training data that is selected by user as relevant and irrelevant. Here, x_i denotes local WCS–LBP feature vector extracted from ith retrieved image and y_i is marked with either 1 (relevant) or 0 (irrelevant) by the user as the class label associated with x_i. From training dataset is constructed, RF is retrained and all candidate images are used as input to the trained random forest. Then the confidence score of each candidate image is updated by using Eq. 5. Equation 5 is the linear combination of previous confidence score and current class distribution $p(c_i \mid L)$ of RF if it is classified to relevant class.

$$Cf_i^{t+1} = \alpha \times Cf_i^{t-1} + (1-\alpha) \times p(c_i|L) \tag{5}$$

In contrast, if a candidate image is classified to irrelevant class, the confidence score of each candidate image is updated by using Eq. 6.

$$Cf_i^{t+1} = \alpha \times Cf_i^{t-1} - (1-\alpha) \times p(c_i|L) \tag{6}$$

In Eqs. 5 and 6, $0<\alpha\leq 1$ is an adjustable parameter to estimate confidence score for next iteration and t labels the iteration time. In this study, the most appropriate α was found to be 0.2 on the basis of several experiments. To determine the proper parameter, seven sample queries (toe, four child nodes of hand, and two child nodes of breast) were tested. In the experiment, we found that among the ten possible values from 0.1 to 1.0, 0.2 produced the best average precision as 93.02%, as shown in Fig. 5. Therefore, adjustable parameter α was adopted as 0.2.

Then, the final distance is re-estimated by sorting the new confidence values and the top k nearest images are displayed as the same method. This process is repeated until all of the desired images are found. The order of the proposed relevance feedback is the following four steps.

1. Let q be the current query keyword. Find all candidate images including the same keyword.
2. Compute k nearest images from candidate images using confidence scores and their descending sorting.
3. User marks N images as relevant or irrelevant.
4. While [user is not satisfied] Do
   (a) Construct $I = \{(x_i, y_i)\}_{i=1}^{N}$ with marked N images.
   (b) Retraining RF using training data in $I = \{(x_i, y_i)\}_{i=1}^{N}$.
   (c) Update confidence scores using Eqs. 5 and 6

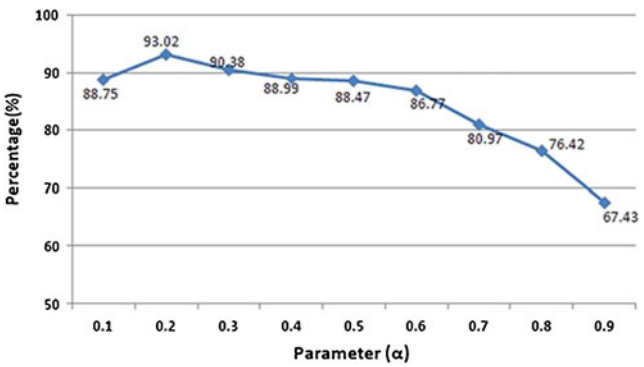

Fig. 5 Experimental results for ten values used to determine adjustable parameter α

(d) Compute k nearest images using new confidence scores.
(e) User marks the N images as relevant or irrelevant.

Figure 6 shows the block diagram of proposed relevance feedback framework.

Experimental results

This system is developed in Visual C++ 2008 language as offline system for training and the test system was developed for on-line based on ASP.NET 3.5 using C# language. The proposed method is applied to our retrieval system, Medical Image Searching System and it can be demonstrated at the following website (http://cvpr.kmu.ac.kr/miss2).

The database images are subset of the ImageCLEF benchmark for image annotation and retrieval as mentioned in the "Introduction" section. To perform the training, 900 images were randomly selected from 30 image categories and each class has equal 30 images. For the test, 1,500 images which did not use for training were used and each class has equal 50 images as shown in Table 1.

First, to measure the annotation performance, we compare the proposed classification method with the widely used MSVM [4–6] with the same local WCS–LBP by estimating the error rate and error count. The error rate means the percentage of codes that have at least one error in one position within one axis (T, D, A, B). The error count gives a greater penalty for misclassification in higher hierarchical instances than for less precise classification in lower hierarchical positions [6, 17].Thus, an image where all positions in all axes are wrong has an error count of 1, and an image where all positions in all axes are correct has an error count of 0. In this paper, because T and B codes are assigned automatically according to the classification result, we only evaluated the error rate and error count on D and A codes.

As can be seen from Fig. 7, the annotation performance of the MSVM with local WCS–LBP shows 25.5% for error rate, 57.31(A) and 36.96(D) for error count. In contrast, RF with the WCS–LBP method showed 20.3% for error rate, 38.8(A) and 23.1(D) for error count, respectively. The main reason of the good performance of proposed methods in error rate and count is that RF was able to to be effective in a large variety of proposed high-dimensional local WCS–LBP, with high accuracy. In addition, RF uses an ensemble of distributions of trees that are trained on only small random subsets of the data and it helps to speed up training and reduce overfitting [20]. In contrast, because MSVMs are not suitable when the feature has high-dimensionality and the database contains over 1,000 images, it gave higher error than RF.

Second, to validate the effectiveness of image-retrieval approach, we compare the retrieval precision using our confidence score assigning method with equal confidence score for all keywords by changing the top k retrieved images. First, three different experts were asked to select 40 representative ground-truth images from each category, and only those images where at least two experts were in agreement were then used for the precision comparison. The test is performed on 30 categories and five queries from each category. In all experiments, performance is measured using average retrieval precision. As shown in Table 2, the overall performance of our approach outperforms the first method regardless of the number of top k as by average percentages of 77% and 71%.

Third, to validate the effectiveness of our feedback approach, we compare the retrieval performance of our algorithm with feature reweighting [2], SVM-based feedback [9] combining confidence score, and our proposed feedback method based on RF using 12 test keywords by increasing the numbers of iteration and top k. Figure 8 shows the

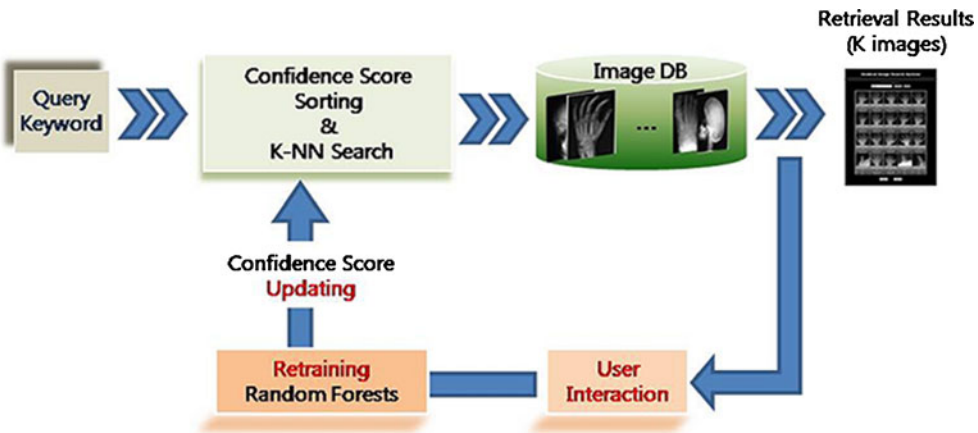

Fig. 6 Block diagram of the proposed relevance feedback framework

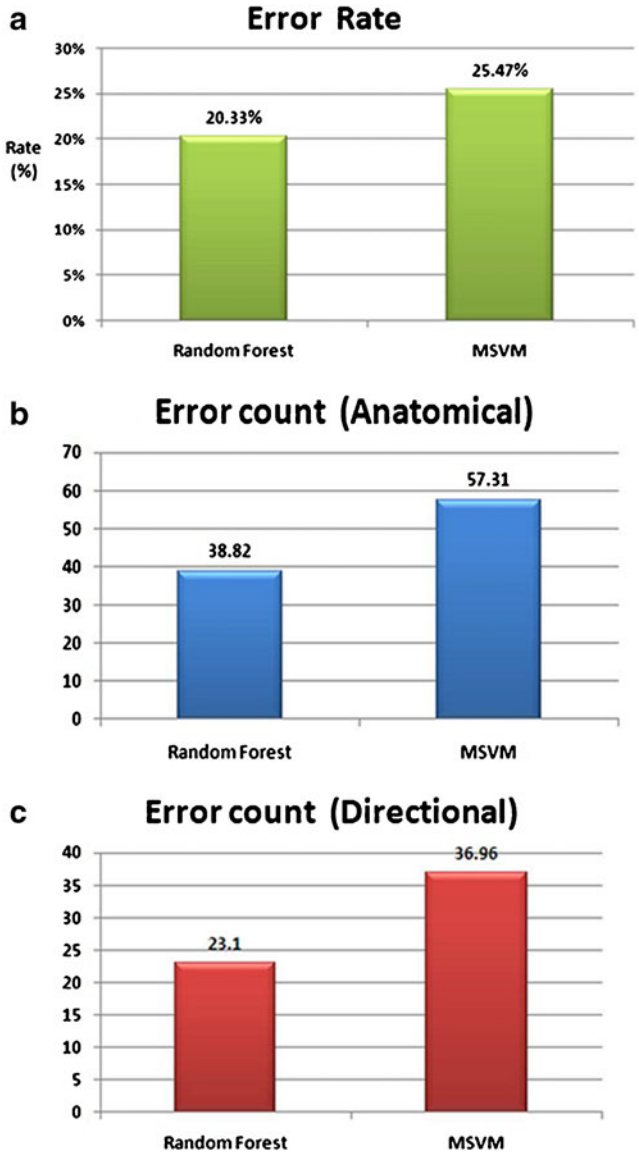

Fig. 7 Comparison of annotation performance between MSVM and the proposed method: **a** comparison of the error rate, **b** comparison of the error count of A code, **c** comparison of the error count of D code

performance comparison for each feedback algorithm evaluated by precision and recall. As we can see from Fig. 8, our approach consistently increases the performance as the amount of iteration time increases. The average precision and recall of the proposed method on four different top *k*s is about 86.2% and 82.6% without relevance feedback. However, after three iterations, the average precision and recall of the proposed method is increased to 96.3% and 90.1%. In contrast, SVM-based feedback showed average precision and recall on four different top *k*s is about 83.5% and 78.1% without relevance feedback. After three iterations, the average precision and recall is increased to 93.2% and 85.7%. Feature reweighting method showed lowest average precision and recall on four different top *k*s is about 62.3% and 61.7% without relevance feedback. After three iterations, the average precision and recall is increased to 68.8% and 68.3%, respectively. Even though SVM-based feedback method shows higher precision and recall rate as 24.4% and 17.4% than reweighting method after three iterations, it has 3.1% and 4.4% lower precision and recall rate than proposed method. The main reasons for the higher precision and recall rate of the proposed and SVM-based feedback method are that two methods not only retrieve similar images using probability and membership values based on image classification results, but also combine the confidence scores with probability and membership values. Moreover, the proposed algorithm is able to retrieve more similar images than SVM-based method regardless of the number of iterations. The main reason for the higher performance of the proposed is that proposed method used RF for image classification and learning feedback classifiers. In particular, RF can construct robust decision trees (classifiers) using only small subsets of the feedback samples and this structure helps to reduce overfitting than SVM.

Figure 9 shows a retrieval result of "Cranium" on top 10 retrieval results without iteration and after first iteration. As shown in Fig. 9, third, sixth, and tenth images including sagittal cranium are removed and coronal cranium images are ranked in higher order after iteration.

Conclusion

In this paper, we proposed three frameworks: multiple keywords annotation for medical images, image retrieval using the keywords and their confidence scores, and relevance feedback for reducing semantic gap between the user and a retrieval system.

Table 2 Comparison of image retrieval performance using two different methods of assigning confidence (percentage)

	Top=10 (%)	Top=20 (%)	Top=30 (%)	Average precision (%)
Equal confidence score	77	70.5	65.4	71
Proposed confidence score	82	76.7	72.2	77

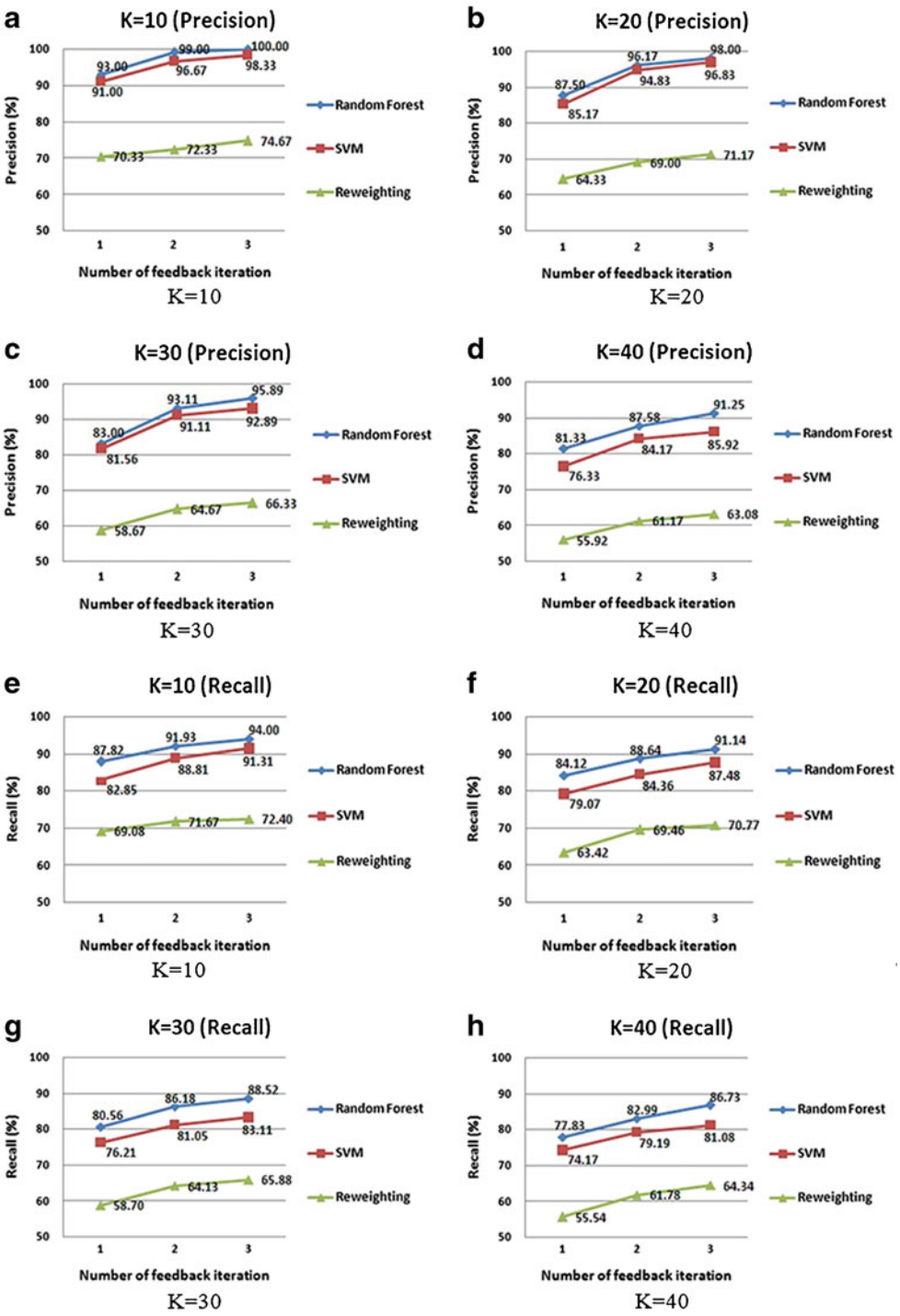

Fig. 8 Comparison of the performance of three learning algorithms—the reweighting algorithm [1], SVM-based algorithm [8], and proposed RF-based algorithm—using five test keywords according to the number of iterations and top *k*: **a**–**d** show the average precision of the three different algorithms according to the increasing top *k*; **e**–**h** show the average recall of the three different algorithms according to the increase in top *k*

To improve keyword annotation performance, this study first proposed a novel medical image classification method combining local WCS–LBP with random forests. Second, for semantic keyword based image retrieval, confidence score was assigned to each annotated keyword by combining probabilities of random forests with predefined body relation graph. To overcome the limitation of keyword-based image retrieval, we combine image retrieval based on keyword and relevance feedback mechanism based on visual feature and pattern classifier.

The experimental results using ImageCLEFmed2007 images showed that the proposed system could indeed improved the annotation performance, average retrieval precision, recall and precision of relevance feedback when each algorithm of our system compared to other methods. In future works, we plan to apply our keyword annotation

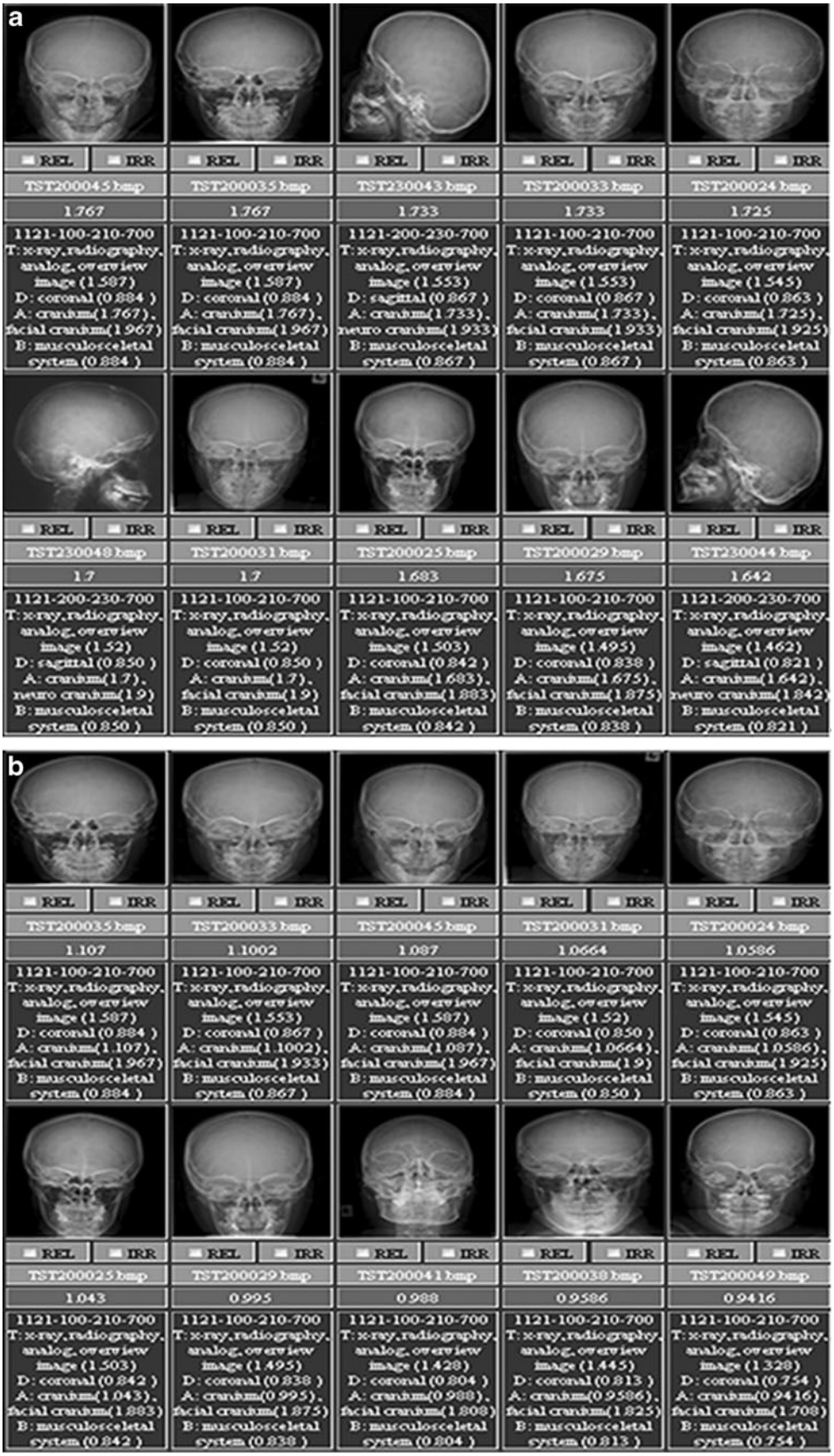

Fig. 9 Example of keyword-based image retrieval and relevance feedback using visual features for the keyword "Facial cranium": **a** initial retrieval results obtained using only the keyword "Facial cranium" and **b** relevance feedback results obtained using visual features and RF after the first iteration

and retrieval algorithm to other medical images, such as cell images, CT images, and MRI images.

References

1. Müller H, Ruch P, Geissbuhler A: Enriching content-based image retrieval with multi-lingual search terms. Swiss Med Inform 54:6–11, 2005
2. Rahman M, Desai BC, Bhattacharya P: Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion. Comput Med Imaging Graph 32:95–108, 2008
3. Julio VR, José GC, José GM, José MF: MIRACLE's naïve approach to medical images annotation. Proceeding of the Workshop on Cross Language Evaluation Forum: 1–9, 2005.
4. Setia L, Teynor A, Halawani A, Burkhardt H: Grayscale medical image annotation using local relational features. Pattern Recognit Lett 29:2039–2045, 2008
5. Mueen A, Zainuddin, Baba MS: Automatic multilevel medical image annotation and retrieval. J Digit Imaging 21:290–295, 2008
6. Amaral IF, Coelho F, Costa JF, Cardoso JS: Hierarchical medical image annotation using SVM-based approaches. Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine: 1–5, 2010
7. Xu X, Lee DJ, Antani SK, Long LR, Archibald JK: Using relevance feedback with short-term memory for content-based spine X-ray image retrieval. Neurocomputing 72:2259–2269, 2009
8. Tong H, He J, Li M, Ma WY, Zhang HJ, Zhang C: Manifold-ranking based keyword propagation for image retrieval. Comput Med Imaging Graph 32:95–108, 2008
9. Bao Y, Zhang Y, Wang D, Shi J: Soft SVM and novel sampling rule based relevance feedback for medical image retrieval. Proceeding of Fourth International Conference on Computer Sciences and Convergence Information Technology: 483–488, 2009.
10. Liu H, Zhang CM, Han H: Medical image retrieval based on semi-supervised learning. J Adv Mater Res 108:201–206, 2010
11. Oh JH, Naqa IE: Adaptive learning for relevance feedback: application to digital mammography. Med Phys 37:4432–4445, 2010
12. Wei CH, Li CT: Learning pathological characteristics from user's relevance feedback for content-based mammogram retrieval. Proceedings of eighth IEEE International Symposium on Multimedia pp: 738–741, 2006.
13. Ko BC, Kim SH, Nam JY: X-ray image classification using random forests with local wavelet-based CS-local binary patterns. J Digit Imaging, 2011. doi:10.1007/s10278-011-9380-3O
14. MacArthur SD, Brodley CE, Shyu CR: Relevance feedback decision trees in content-based image retrieval. Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries: 68–73, 2000.
15. Lakdashti A, Ajorloo H: Content-based image retrieval based on relevance feedback and reinforcement learning for medical images. ETRI J 33:240–250, 2011
16. ImageCLEF. Available at http://www.imageclef.org/2007/photo. Accessed 25 April 2011.
17. Tommasi T, Orabona F, Caputo B: An SVM confidence-based approach to medical image annotation. Lect Notes Comp Sci 5706:696–703, 2009
18. Ojala T, Pietikainen M, Maenpaa T: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24:971–987, 2002
19. Heikkilä M, Pietikäinen M, Schmid C: Description of interest regions with local binary patterns. Pattern Recognit 42:425–436, 2009
20. Breiman L: Random forests. Mach Leaning 45:5–32, 2001
21. Heidemann G: Unsupervised image categorization. Image Vision Comput 23:861–876, 2005